2.5 Spatial Blurring Features

Spatial resolution degradation is an artifact that normally occurs when a video camera is imperfectly focused or when motion is present in a video scene. Camera defocusing reduces spatial resolution by spreading incident light over a larger surface area. Thus, a defocused camera is unable to pass the high spatial frequency information present in imagery containing sharply defined edges and fine detail. Under conditions of video motion, the bandwidth compression techniques typically employed in digital video systems are unable to retain enough of the high frequency information to avoid blurring of the edges. Investigators in the fields of human vision and human object recognition have recognized the importance of sharp edges for correct visual perception and recognition of objects (Shapley and Tolhurst, 1973, Held et al., 1978, Geuen and Preuth, 1982, Beiderman, 1985, Owens et al., 1989). The importance of sharp edges for moving objects is currently a research topic and appears to depend on whether or not the eye can track the object. Several methods have been proposed to detect automatically the sharpness of image edges. In section 2.5.1, the procedure of Toit and Lourens (1988) for estimating the edge sharpness of arbitrary video imagery has been adapted to measuring the spatial resolution degradation present in digitally transmitted video systems.

2.5.1 Feature Extraction Technique

A method for estimating the sharpness of edges in sampled video imagery can be obtained using very simple image processing techniques. The procedure relies on being able to sample the input (undistorted) video imagery as well as the output (distorted) video imagery. The input video imagery is required as a reference so that the amount of spatial resolution degradation present in the output imagery can be estimated. The edge sharpness feature can be extracted by computing the amount of energy present in the edge extracted video imagery. The theory is that sharper edges will contribute more high intensity pixel values than blurred edges. Several steps are required to apply the technique:

1. Video alignment

   If one desires to observe the instantaneous value of the feature (frame-by-frame), then single-frame temporal alignment of the input and output video is recommended. Strictly speaking, time alignment of the input and output video is not required to extract this feature, provided one only requires the average value of the feature (over all frames in the video sequence).

2. Video preconditioning

The sampled video imagery is preconditioned to remove edge energy contributions resulting from camera interlace effects and noise spikes. Because edge extraction filters (to be applied in step 3 below) involve taking the difference of neighboring image pixel values, they will enhance noise in the imagery as well as edges. Therefore, some preconditioning of the imagery is strongly desired before application of the edge extraction filter.

If noise spikes are present, they can be removed by use of a small median filter that does not significantly blur the edges (Tzafestas, 1986, Gonzalez and Wintz, 1987, Jain, 1989). Fine detail, such as object corners, will be blurred by the median filter. For the imagery to be presented later in this report, median filtering was performed (see Appendix B for a description of the median filter that was used).

Camera interlace effects may be present in an NTSC video system when the video scene contains objects that are moving horizontally. In Figure 4, vertical edges are seen to form many alternate horizontal edges that have a length proportional to the velocity of horizontal motion. The horizontal edges caused by camera interlace will contribute a large amount of erroneous edge energy. Here, the interlace effects could be removed by sub-sampling the image by a factor of 2 in both horizontal and vertical directions (every other row and column in the image being discarded). A more desirable method of reducing the erroneous edge energy due to interlace would be to select an edge extraction filter which is insensitive to interlace. This is the recommended method and the one which is used here in step 3 below. The Sobel edge extraction filter (Tzafestas, 1986, Gonzalez and Wintz, 1987, Jain, 1989) is insensitive to interlace effects because it detects edges by computing the difference between image pixel values that are spaced two pixels apart (see Appendix B for a description of the Sobel filter). Thus, the Sobel filter does not extract the edges due to alternating black and white interlace lines (as shown in Figure 4).

(a) Slow Motion

(b) Fast Motion

Figure 4. Camera interlace effects caused by horizontal motion.

3. Edge extraction

An edge extraction filter is applied to the preconditioned video imagery. The reader is referred to Gonzalez and Wintz (1987) or Jain (1989) for a description of several of the many different types of edge extraction filters. For the imagery to be presented later in this report, a Sobel filter was chosen.

4. Feature computation

Several features can be computed from the edge extracted or Sobel filtered imagery. Four are suggested here:

a. The mean of the Sobel image (M-SI)

M-SI is computed as the summation of the image pixel values divided by the total number of pixels. Here, the summation can be performed over any sub-regional area of the image. See Appendix A, equation 3 for a mathematical definition of M-SI.

b. The standard deviation of the Sobel image (SD-SI)

SD-SI is computed as the square root of (the summation of the squares of the image pixel values divided by the total number of pixels, minus the square of M-SI). This estimate of the standard deviation is asymptotically unbiased for a large number of image pixels, which is typically the case. See Appendix A, equation 4 for a mathematical definition of SD-SI.

c. The root mean square of the Sobel image (RMS-SI)

RMS-SI is computed as the square root of (the summation of the squares of the image pixel values divided by the total number of pixels). See Appendix A, equation 5 for a mathematical definition of RMS-SI.

d. The number of pixels greater than a threshold of the Sobel image (NPGT-SI)

NPGT-SI is computed as the total number of pixels within any sub-regional area that exceed a fixed threshold. Advantages of this feature include the ability to detect the blurring of just the sharpest edges, and ease of computation. There is some indication that humans may perform quality assessment by examining the sharpest high contrast edges (Westernik and Roufs, 1988). The higher the threshold for NPGT-SI, the sharper the edges must be before being included in the summation. Subjectively judged video data could be used to determine the threshold setting that gives the best correspondence with subjective quality. Since subjective data was unavailable at the time of this report, a somewhat arbitrary threshold was selected that included the predominate edges of the image. See Appendix A, equation 6 for a mathematical definition of NPGT-SI.

The decrease in the amount of edge energy that the output imagery has with respect to the input imagery can be used to estimate the amount of spatial resolution degradation present in the output imagery with respect to the input imagery. Alternately, since the input imagery contains sharper edges (and hence higher pixel values) than the output imagery, the decrease in the number of pixel values that exceed the threshold can be used to estimate the spatial resolution degradation. In either case, by normalizing with the reference imagery, a feature can be formed that varies between 1 (no edge blurring) to 0 (complete edge blurring).

The features described above exhibit many of the desirable properties of features mentioned earlier. In particular, the features are applicable to arbitrary video scenes and may be applied to any sub-region of the image, making them useful for local estimates of image quality. In addition, the feature extraction process is computationally efficient and stable. The features are insensitive to noise spikes (due to median filtering) and small gray level shifts in the image (since edges are computed from the differences of neighboring image pixel values and thus, the background gray level subtracts off).

2.5.2 Sample VTC/VT Results

For illustrative purposes, the method for extracting the spatial blurring features was applied to several VTC/VT video frames sampled by an 8 bit video frame grabber. Figure 5 shows the sampled imagery after median filtering. The top left image was captured from an NTSC camera with no motion present in the video scene. The top right image was captured from an NTSC camera when rotational motion was present in the video scene. The bottom right image was captured from the output of a VTC/VT codec running at the digital signal 1 (DS1) rate of 1.544 Mbps, where the input imagery to the codec was the same as that shown in the

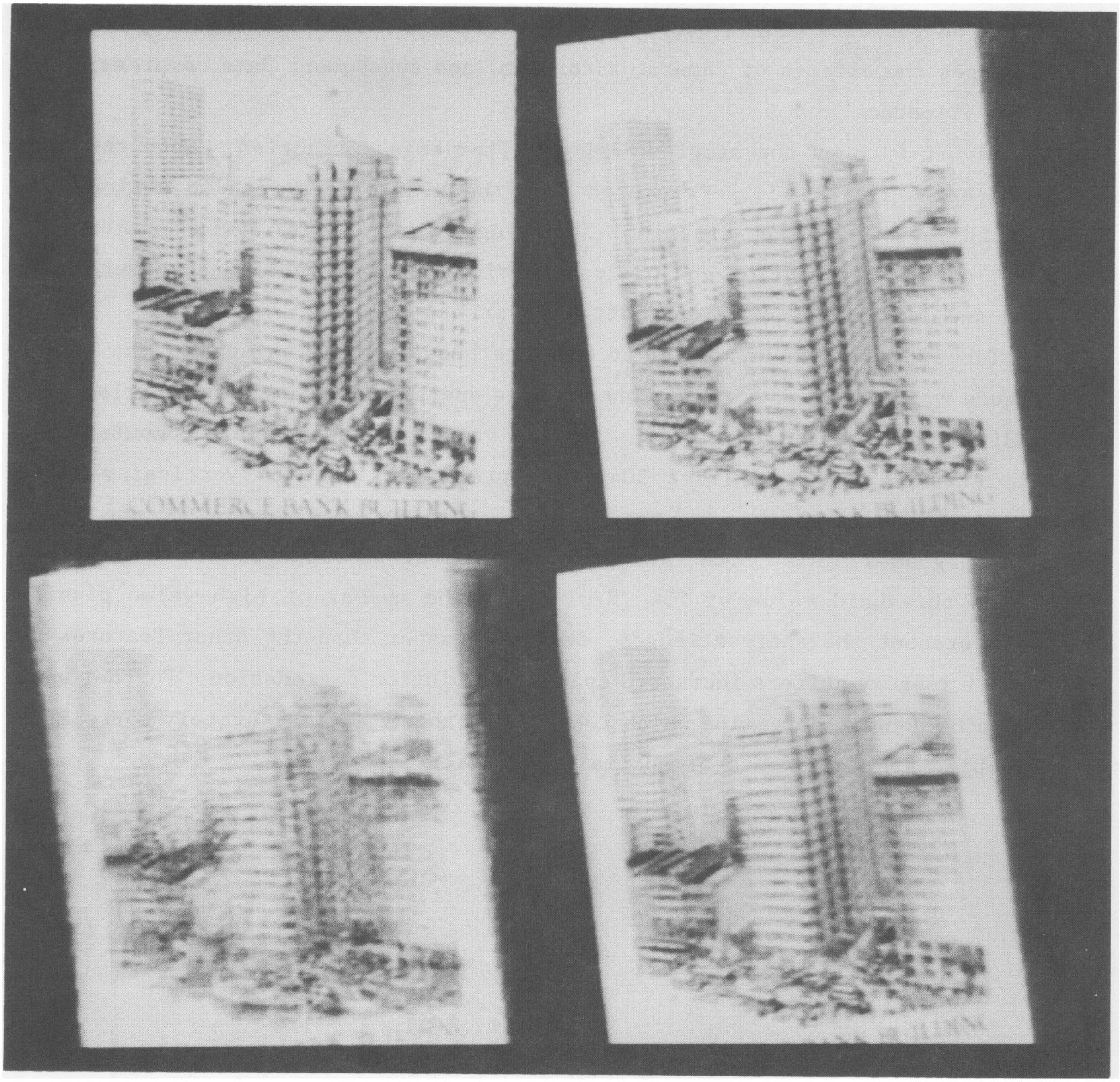

Figure 5. VTC/VT imagery containing rotational motion. Top left – camera with no motion. Top right – camera with rotational motion. Bottom right – camera with motion and DS1 data compression. Bottom left – camera with motion and 1/4 DS1 data compression.

top right image of Figure 5. The bottom left image was captured from the output of a VTC/VT codec running at rate 1/4 DS1. In Figure 5, one can clearly see the effects of camera distortion, and subsequent data compression by the VTC/VT codec.

Figure 6 shows the sampled imagery after edge extraction. Note the well defined edges for the image captured from the NTSC camera with no motion (top left) and the successive worsening of the edge blur for camera with motion (top right), camera with motion and DS1 compression (bottom right), and camera with motion and 1/4 DS1 compression (bottom left).

Table 2 shows the unnormalized edge sharpness feature values for the images in Figure 6. To eliminate the erroneous edge energy at the image boundaries (due to median and Sobel filtering), the feature values in Table 2 were computed over a sub-rectangular region (size 504 horizontal pixels by 464 vertical pixels) centered on the main image. Note the decrease in edge energy as the image quality degrades. Also note the decreasing number of image pixels that exceed a chosen threshold value of 250 (NPGT-SI). The number of high value pixels, which represent the sharpest edges, decrease faster than the other features as the input imagery suffers increased spatial resolution degradation. Further work needs to be done to determine which feature in Table 2 most accurately correlates with subjective judgements of spatial resolution.

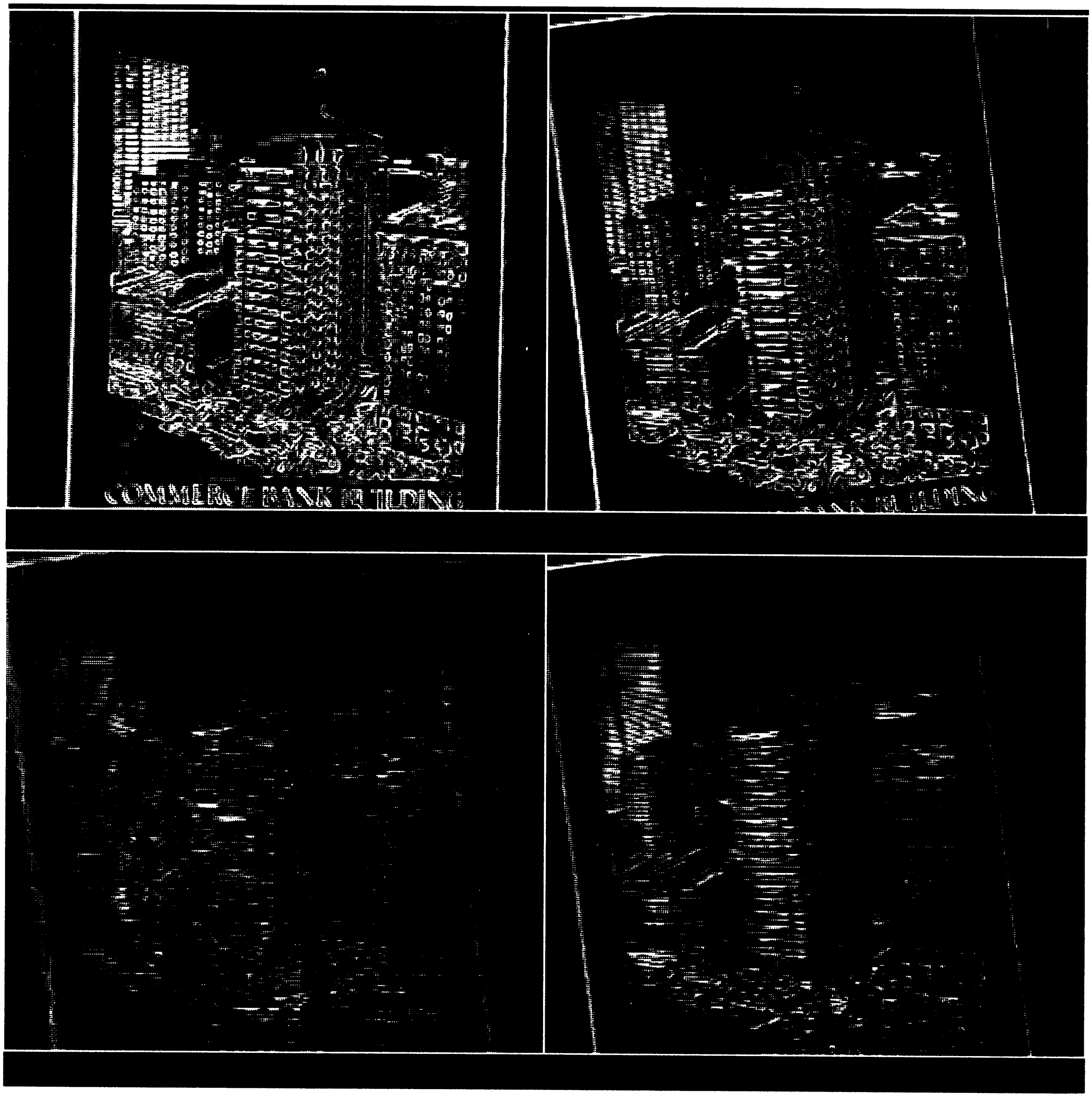

Figure 6. Sobel filtered edge extracted VTC/VT imagery of Figure 5. Note the well defined edges for the image captured from an NTSC camera with no motion (top left) and the successive worsening of the edge blur for camera with motion (top right), camera with motion and DS1 compression (bottom right), and camera with motion and 1/4 DS1 compression (bottom left).

Table 2. Spatial Blurring Features For VTC/VT Imagery Of Figure 6

Image	**M-SI**	**SD-SI**	**RMS-SI**	**NPGT-SI**
Top Left (Still)	59.6	81.2	100.8	9116
Top Right (Camera + rotation)	48.2	64.7	80.7	3882
Bottom Right (Camera + rotation + DS1)	37.3	48.6	61.3	995
Bottom Left (Camera + rotation + 1/4 DS1)	32.2	36.3	48.6	184

The edge sharpness features were also computed for a typical VTC/VT scene that contained upper body motion. The video scene in Figure 7 contains motion of the man's head and hands, and the report that the man is holding. The top row shows a sample of 4 consecutive images that were frame-grabbed from the original NTSC VTC/VT scene. The images were grabbed at each field increment of the video recorder. Thus, the time difference between consecutive images in a row is approximately 1/60 of a second. The second, third, and fourth rows of images were obtained from the output of a VTC/VT codec that compressed the NTSC video to bit rates of DS1, 1/2 DS1, and 1/4 DS1, respectively. The single-frame temporal alignment method has been used to align the codec output video shown in Figure 7. For clarity, the images in the first column (leftmost) of Figure 7 have been expanded in Figure 8. In Figure 8, the top left image is the original NTSC, the top right is DS1, the bottom right is 1/2 DS1, and the bottom left is 1/4 DS1. Note that most of the image distortion occurs locally in areas that contain motion (man's right hand), and that the static background is relatively distortion free. As the codec is forced to operate at lower bit rates, areas of the image that contain motion become more and more blurred.

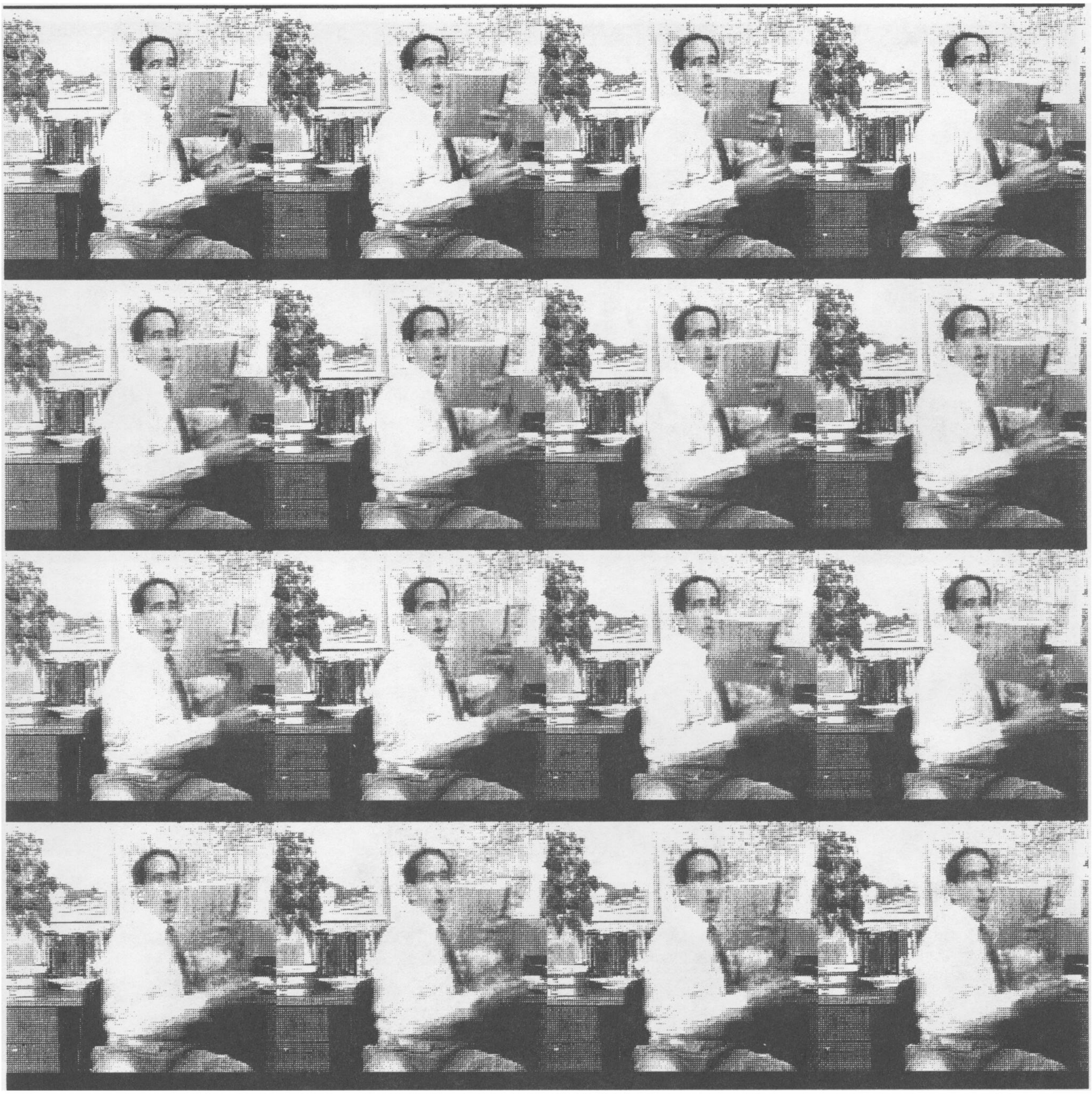

Figure 7. VTC/VT imagery containing upper body motion. Top row - NTSC input. Second row -codec output at rate DS1. Third row - codec output at rate 1/2 DS1. Bottom row - codec output at rate 1/4 DS1.

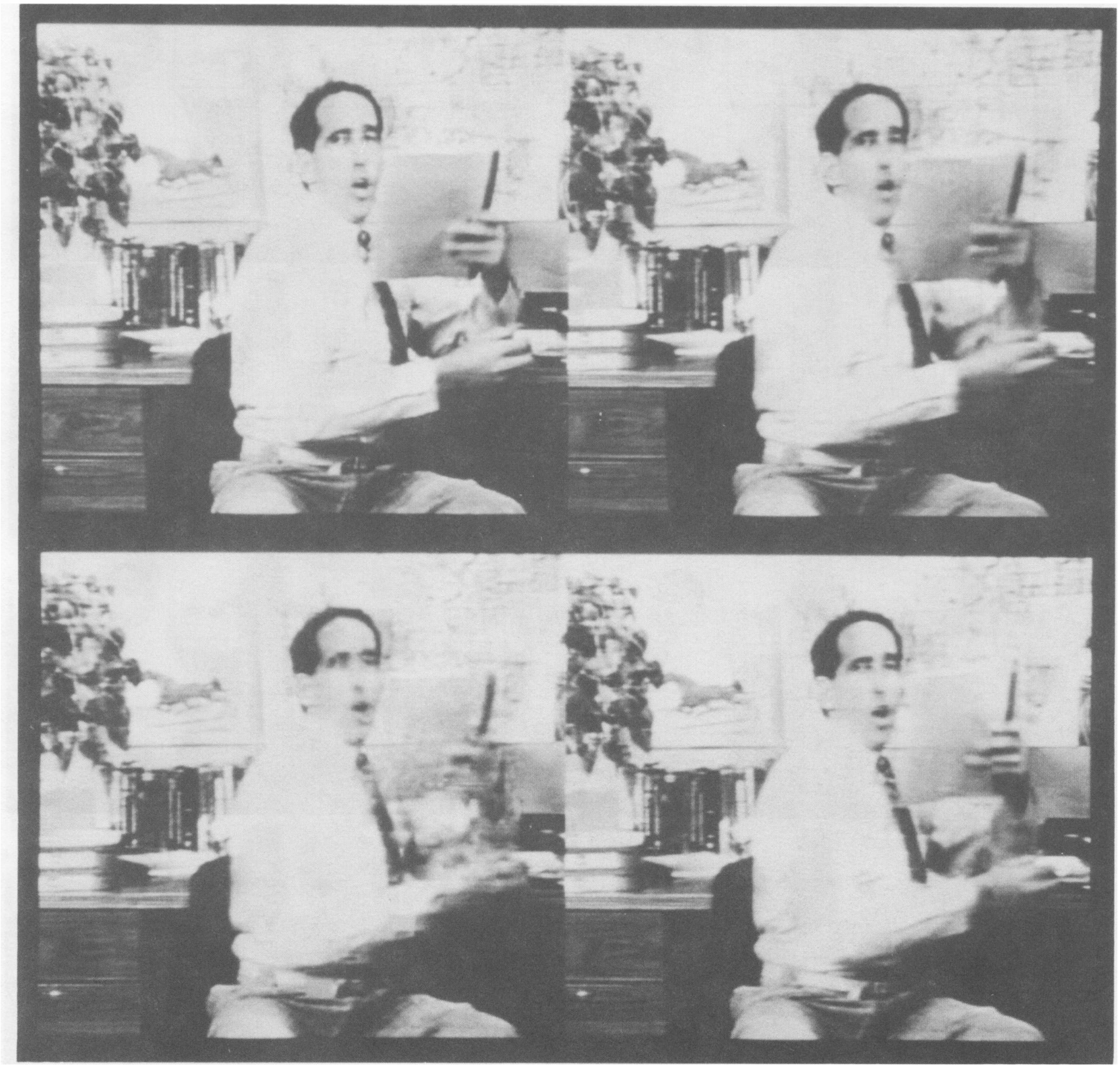

Figure 8. Leftmost column of Figure 7 expanded. Top left - NTSC input. Top right - codec output at rate DS1. Bottom right - codec output at rate 1/2 DS1. Bottom left - codec output at rate 1/4 DS1.

Figure 9 shows the video of Figure 7 after median filtering and edge extraction. For clarity, Figure 10 shows the expanded video of Figure 8 after median filtering and edge extraction. Note that edges of moving objects appear less intense as the codec is forced to operate at lower bit rates. Thus the edges are most blurred for images in the bottom row (bit rate of 1/4 DS1). Table 3 shows the average of the unnormalized spatial blurring features for eight consecutive images, the first four of which are shown in Figure 9. The sub-rectangular image regions and thresholds (for NPGT-SI) that were used to generate Table 2 were also used to generate Table 3. The features in Table 2 were generated from video imagery that contained rotational motion which included a large part of the image. The features in Table 3 were generated from video imagery that contained only a small amount of natural motion. The codec performs differently for the two types of video scenes, and this is reflected in the computed features.

Since subjective quality ratings are based on a video scene that is normally 10 seconds long, a very robust process would be to extract the features from many frames of video, and even from many sub-regions of each video frame. Then, the feature classification system (shown in Figure 1) could utilize all of the feature samples to improve the video quality classification.

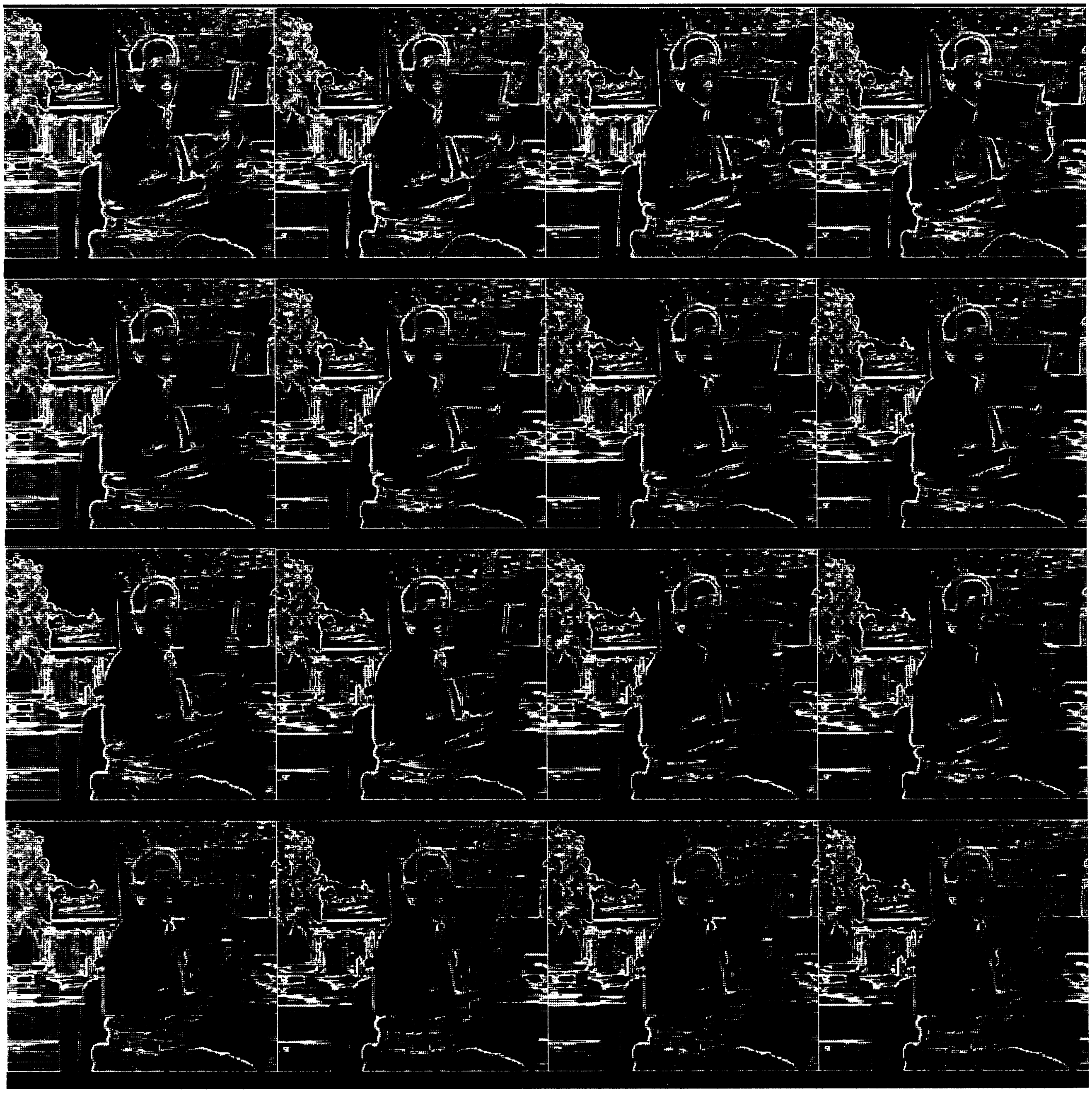

Figure 9. Sobel filtered edge extracted VTC/VT imagery of Figure 7. Note the well defined upper body edges for the images captured from the NTSC input (top row) and the successive worsening of the edge blur for DS1 compression (second row), 1/2 DS1 compression (third row), and 1/4 DS1 compression (bottom row).

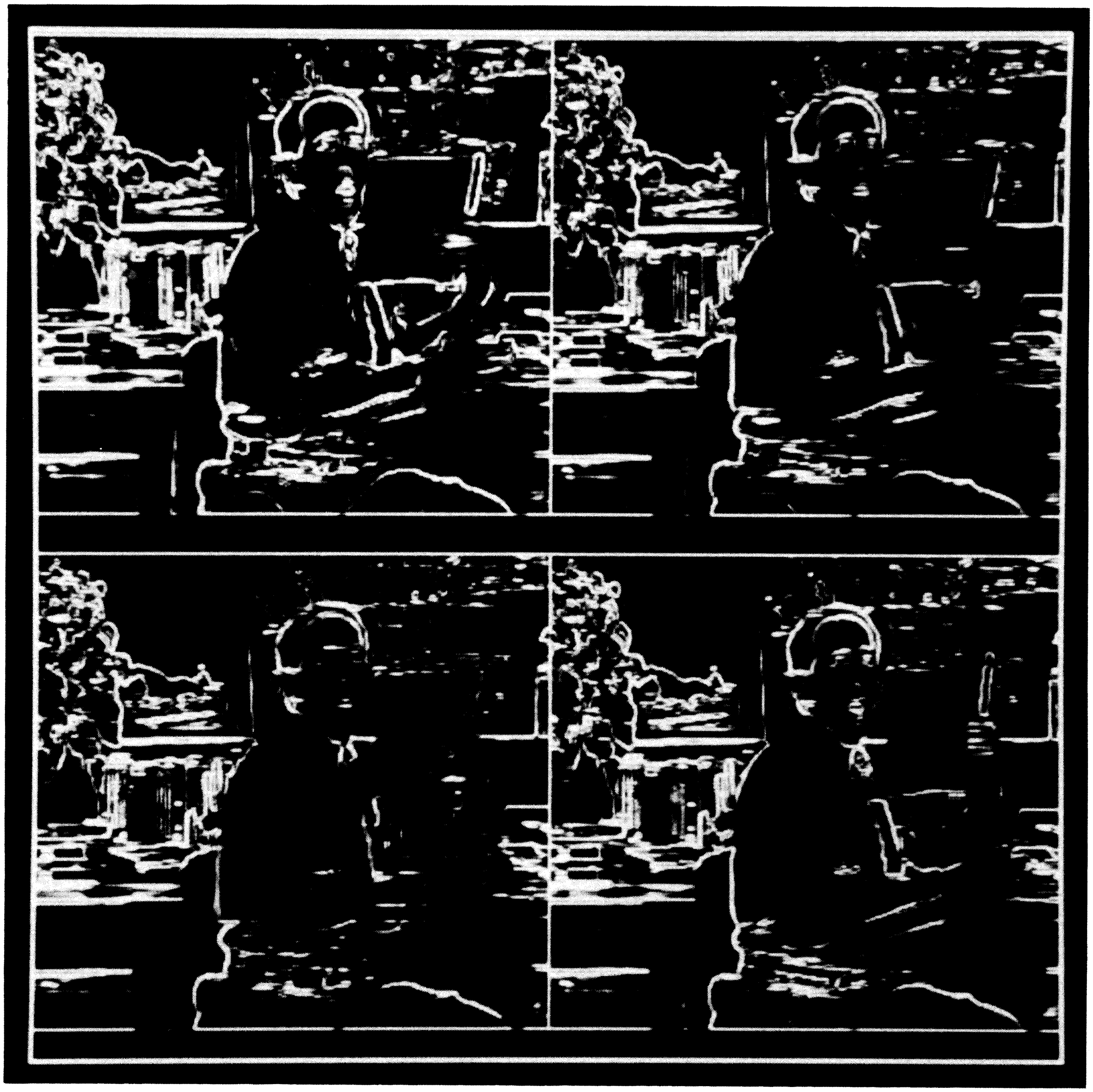

Figure 10. Sobel filtered edge extracted VTC/VT imagery of Figure 8. Note the well defined upper body edges for the images captured from the NTSC input (top left) and the successive worsening of the edge blur for DS1 compression (top right), 1/2 DS1 compression (bottom right), and 1/4 DS1 compression (bottom left).

Table 3. Spatial Blurring Features For VTC/VT Imagery Of Figure 9

Scene	M-SI	SD-SI	RMS-SI	NPGT-SI
Top Row (NTSC)	70.4	103.3	125.1	14388
Second Row (DS1)	61.7	90.0	109.1	10247
Third Row (1/2 DS1)	60.9	89.5	108.3	9972
Bottom Row (1/4 DS1)	59.2	83.8	102.6	8265

2.6 Blocking, Edge Busyness, and Image Persistence Features

Blocking, defined in Table 1, is a severe form of spatial resolution degradation that normally occurs at low codec bit rates when there is a lot of motion in some sub-region or all of the video scene (such as during camera pans or zooms). Edge busyness and image persistence, also defined in Table 1, are video coding artifacts that causes false activity to appear around edges or elsewhere is the video scene. Blocking, edge busyness, and image persistence are most noticeable when the motion involves a high contrast (sharp) edge. Blocking, edge busyness, and image persistence cause edge energy to appear in the output video scene that was not present in the original input video scene. Human viewers have semantic knowledge of how certain items should look and they take objection to the presence of erroneous, out of place artifacts such as blocking, edge busyness, and image persistence. In particular, the appearance of false regular edge energy such as blocking is very noticeable and objectionable to the human viewer (more so than spatial blurring). Therefore, it is desirable to have a set of features that only measures the amount of false edge energy in a video scene.

Section 2.6.1 proposes a technique for extracting a set of features that quantitatively measures the amount of false edge energy in the output video scene. The features may be used to measure blocking, edge busyness, and image persistence since all contribute false edge energy to the output video scene. A by-product of the false edge energy feature is another set of features for measuring spatial blurring. The new